

INTERNSHIP AT GKN AEROSPACE

Efficient Data Mining and Information Extraction

Shreya Manepalli

Dept. of Computer Science, PES University, Bangalore

TABLE OF CONTENTS

1. Purpose

2. Abstract

3. Introduction

3.1 Data, structured and unstructured

3.2 Text mining and its importance

4. Project Overview

4.1 Non-Conformance Reports (NCRs) Data Extraction

4.1.1 Non-Conformance Reports

4.1.2 Project Importance

4.2 Color Codes Extraction from Images for the Fokker Plant

4.2.1 Project Importance

4.3 Generalized Data Mining from PDF and Word Documents

4.3.1 Project Importance

5. Objective

6. Methodology

7. Results

8. Testing and Performance

9. Conclusion

10. Future scope

10.1 Non-Conformance Reports (NCRs) Data Extraction

10.2 Color Codes Extraction from Images for the Fokker Plant

10.3 Generalized Data Mining from PDF and Word Documents

1. Purpose

The purpose of this project report is to document the outcomes and achievements of my internship at GKN Aerospace, where I worked on data mining and efficient information extraction from data sources such as pdf, word files and images. Over the course of the internship, I undertook three distinct projects, each contributing to the overall objective of leveraging data for valuable insights and analytics.

The overall aim of this project report is to showcase the value and potential of data mining and information extraction from unstructured sources in a real-world industrial context. It will serve as a valuable record of the work accomplished during the two-month internship at GKN Aerospace and contribute to the knowledge base for future data-driven endeavors.

2. Abstract

GKN Aerospace generates a substantial volume of unstructured data annually, comprising reports, documents, and images. This internship aimed to leverage data mining techniques to extract valuable information from this vast repository of PDFs, word files, and images enabling the creation of insightful datasets for data analysis using Python.

This work encompasses three distinct projects. In the first project, I extracted specific data from Non-Conformance Reports (NCRs) and organized it into an Excel file. This curated dataset can now be used for data analytics. The second project involved extracting color codes from images of airplane wings and systematically storing them in a structured database. The third project revolved around generalized data mining from PDF and Word documents. These extraction algorithms are designed to efficiently capture various data elements, including tables, equations, relevant text, figure/table captions, and data from images. The Python codes developed can be adopted for similar usage with minimum changes.

3. Introduction

3.1 Data, structured and unstructured

Data refers to any collection of facts, statistics, measurements, observations, or information that can be recorded, stored, and analyzed. In the context of computing and technology, data is essential for various purposes, including decision-making, analysis, and generating insights. Data can take many forms, such as text, numbers, images, audio, video, and more.

Structured Data:

Structured data refers to information that is organized and formatted in a specific and predefined manner. It follows a fixed schema or data model, where each data item has a well-defined data type and is organized into rows and columns, similar to a table in a database. The schema provides a clear representation of the data, making it easy to understand and work with. Examples of structured data include data in relational databases, spreadsheets, and organized data in CSV files.

Unstructured Data:

Unstructured data refers to information that does not follow a predefined structure or format. It does not fit into traditional databases with rows and columns, and its organization and content may vary widely. Analyzing and extracting meaningful insights from unstructured data require specialized tools and techniques, such as natural language processing (NLP) for text data and computer vision for image data. Examples of unstructured data include text documents, images, audio files, social media posts, videos, emails, and web pages.

3.2 Text Mining and its Importance

Text mining, also known as text data mining or text analytics, is the process of extracting useful information and knowledge from large volumes of unstructured text data. It involves using various techniques from natural language processing (NLP), machine learning, and statistical analysis to analyze, interpret, and derive meaningful insights from textual data. Text mining is crucial in unlocking the hidden value within unstructured text and is important in various fields and industries.

4. Project Overview

4.1 Non-Conformance Reports (NCRs) Data Extraction

In the first project, I focused on extracting specific data from Non-Conformance Reports (NCRs) and organizing it into a structured Excel file. This curated dataset now serves as a valuable resource for data analytics and can be utilized to gain deeper insights into various products and processes.

4.1.1 Non-Conformance Reports

Non-Conformance Reports (NCRs) are crucial documents that record instances of non-compliance with quality standards or deviations from specified requirements.

4.1.2 Project Importance

Manually reading and entering data from NCRs into Excel can be time-consuming and error-prone. By developing automated data extraction algorithms, I significantly reduced the manual effort involved, leading to increased efficiency and accuracy in data processing.

The curated dataset created from the extracted NCR data now serves as a valuable resource for data analytics and decision-making. Analysts and stakeholders can utilize this structured data to gain deeper insights into various products and processes, identify trends, and make data-driven improvements.

4.2 Color Codes Extraction from Images for the Fokker Plant

The second project involved developing algorithms to extract color codes from images and systematically storing them in a structured database.

4.2.1 Project Importance

By automating the extraction of color information from images, I streamlined the processes that otherwise would have required manual identification and recording of color codes. This automation not only saves time but also reduces the chances of human errors.

4.3 Generalized Data Mining from PDF and Word Documents

The third project centered around the development of versatile extraction algorithms capable of efficiently capturing various data elements from PDF and/or Word documents. These elements include tables, equations, relevant text, figure/table captions, and data from images. The Python codes developed in this project can be adapted for similar data mining tasks with minimal modifications, increasing the overall efficiency of information extraction.

4.3.1 Project Importance

Manually extracting data elements such as tables, equations, relevant text, and captions from documents can be time-consuming. These generalized data mining algorithms have significantly reduced the time and effort required for these tasks, leading to improved efficiency in handling large volumes of data.

Analysts and stakeholders can access relevant information quickly, aiding in research, product development, compliance monitoring, and more.

5. Objective

The primary objective of this internship was to develop robust data mining and information extraction techniques to handle the vast volume of unstructured data generated at GKN Aerospace. By addressing the specific needs of various stakeholders, these techniques aimed to streamline data processing, enhance data analytics capabilities, and ultimately enable data-driven decision-making. The three distinct projects catered to the requirements of different stakeholder groups within the organization:

Project 1: Non-Conformance Reports (NCRs) Data Extraction for the Shared Product Engineering (SPE) Team

Stakeholder: The SPE Team at GKN Aerospace, responsible for analyzing a large number of Non-Conformance Reports (NCRs) that are generated.

Objective: The primary objective of this project was to develop a Python script to extract specific data from Non-Conformance Reports (NCRs) generated and analyzed by the SPE Team at GKN Aerospace. The script aims to automate the data extraction process, as the NCRs contain unstructured data which used to be manually retrieved for analysis. The extracted data will be organized into an excel file, making it suitable for data analytics.

Project 2: Color Codes Extraction from Images for the Fokker Plant

Stakeholder: The Fokker Plant, focusing on airplane wings and their associated color codes.

Objective: The objective of the second project is extract colors and associated codes from the pdf's containing images about airplane wings. These color codes and numbers are present in specific formats within the images and can be stored systematically into Excel files which can be used for further data analysis. This project aimed to ease the work of the operator of manually mining the data and assist the Fokker Plant in quickly accessing and utilizing color code information, thereby enhancing their productivity and accuracy in handling data-specific tasks.

Project 3: Generalized Data Mining from PDF and Word Documents for the GAI Team

Stakeholder: The entire GAI Team at GKN Aerospace, encompassing various departments and individuals with diverse data mining needs.

Objective: To create generalized data mining algorithms capable of efficiently capturing diverse data elements such as tables, equations, relevant text, figure/table captions, and images from PDF and Word documents. By providing a versatile solution that can be adopted by anyone within the organization, this project aimed to empower the GAI Team with a powerful toolset for data extraction, facilitating a wide range of data mining applications.

Overall, the main objective was to deliver practical and adaptable data mining solutions that catered to the specific requirements of the SPE Team, the Fokker Plant, and the entire GAI Team. By automating data extraction processes and structuring the extracted information into databases, the internship sought to decrease the manual effort of operators maximize data utilization, and data analysis, improve productivity and foster data-driven insights across different departments and stakeholders at GKN Aerospace.

6. Methodology

Project1: Data Extraction from Non-Conformance Reports (NCRs)

Data Collection and Preprocessing

The Python script accepts user input to determine whether the input is a single file or a folder containing multiple NCR PDF files. If the input is a single file, the script performs data extraction directly on that file. If it is a folder, the script iterates through all the PDF files present in the folder.

Modules Used

The script utilizes the "tabula" library to extract tabular data from the NCR PDF files. For each PDF file, the script reads all tables on every page and stores them in separate sheets within an intermediate Excel file. Additionally, the script extracts specific information from the first page of the PDF files, including NCR numbers, material numbers, and serial numbers affected, and stores them in memory for later use.

Data Transformation and Organization

The extracted data from each PDF file is organized into separate sheets within the Excel file based on the drawing number associated with each report. This segregation facilitates easy access and analysis of relevant information.

Data Validation and Cleaning

To ensure the accuracy and consistency of the extracted data, the script checks for valid sheet names in Excel, and for any missing data in the NCR's and handles them appropriately.

Working of the code

The script appends the extracted data from each PDF file into the corresponding sheets in the output Excel file. For each NCR, the script appends the item number, description of non-conformance, character number, defect type, measured value, and deviation value.

Project 2: Color Codes Extraction from Images for the Fokker Plant

Data Collection and Preprocessing

The Python script first takes the main parent folder containing PDF files as the input. It utilizes the "fitz" library to convert each PDF file into images with a specific DPI (dots per inch). The converted images are stored in an intermediate output folder, preserving the same folder structure as the original PDF files.

The images are then chunked into smaller regions to enhance readability and to isolate specific information. A function is developed to crop each image into smaller chunks, ensuring each chunk contains meaningful data. The chunking process involves defining the chunk size, adding padding for better separation, and checking for significant data in each chunk using edge detection and contours analysis.

Extracting Colors and their associated codes

The Python script uses the "easyocr" library to read text from the cropped image chunks. For each chunk, the script processes the text to identify color codes and their associated numbers. The script stores the extracted color codes and numbers in a dictionary, where the color codes act as column headers, and the numbers are the corresponding values.

Working of the code

The Python script generates an Excel file for each PDF file processed, containing extracted color codes and numbers. The Excel file is saved in the final output folder, named after the original PDF file.

Optimization

Threading for Parallel Processing is used to improve efficiency, the script utilizes threading to process multiple image chunks simultaneously. The "concurrent. Futures" module enables parallel execution, reducing the overall processing time.

Project 3: Generalized Data Mining from PDF and Word Documents

Data Collection and Preprocessing

The Python script takes pdf or word files as an input and extract various data elements such as tables, equations, relevant text, figure/table captions, and data from images.

Modules used

The tool is built using Python and relies on several libraries for different functionalities, such as PyPDF2 and fitz for PDF processing, docx2txt and docx for Word document processing, tabula and pandas for table extraction, and OpenPyXL for Excel manipulation.

Working

The GUI takes pdf's and word files as an input. The tables from each file are extracted into excel files which can be used for further analysis. The user can enter multiple words or phrases and text relevant to that keyword is extracted into a document. Similarly there is an option to extract all the equations, figure captions, table captions, and images from the files provided by the user as an input.

7. Results

Project 1: Data Extraction from Non-Conformance Reports (NCRs)

The final output of the script is a consolidated Excel file containing organized data from all NCRs processed during the operation. The output file can be used for further data analytics and insights, enabling more efficient decision-making and process improvements.

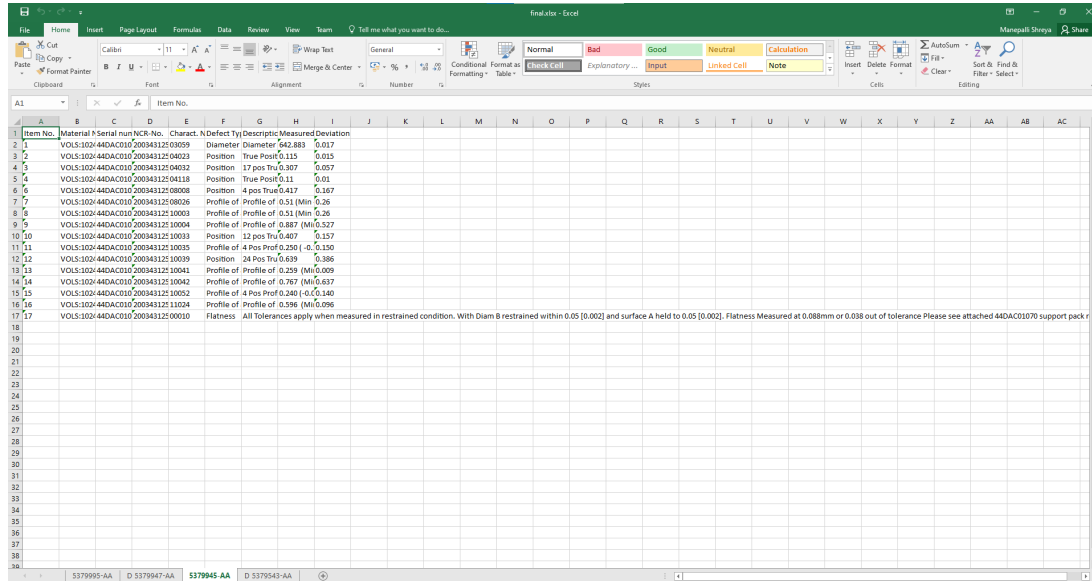
Input given by the user (Sample NCR Report)

3 (15)

Nonconformance report			NCR No.
Material No. VOLIS 10257700	Ver AD	Material Description Fire Containment Assy	200351871
Program PW19000 Embraer Firewall	Design Organization Drawing & issue D:5379543-AA		Total Quantity Order No 1
Coordinator Ann Charlott Linusson	Design Organization Part number & version D:5379543-AA		Reference No. 2023-175

Item No. 2			
Defect Class	Pieces affected 1	Defect type	Character No 04192
Requirement Location PAGE-04	Zone K3		
Serial numbers affected 44DAC02099			
Description of Nonconformance True Position (D:3A/B/C) is 0.329 or 0.029 out of tolerance. Please see attached 44DAC02099SupportPack for further details.			
Cause of Nonconformance For detail Root cause Analysis please refer to 8D investigation #RCA200328488.			
Corrective Action (to avoid re-occurrence of this Nonconformance) Planned from: 2022-03-11 For detail corrective action / preventative action please refer to 8D investigation #RCA200328488.			

Final Output Generated

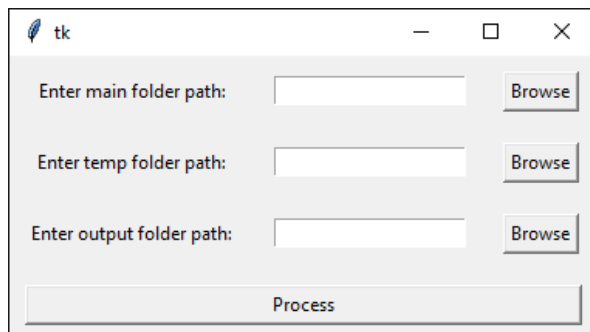


Item No.	Material	Serial num	NCR	No.	Charact	N Defect Ty	Descriptive	Measured	Deviation
1	VOLS-1024-44DAC010	200343125	08059		Diameter	Diameter	542.883	0.017	
2	VOLS-1024-44DAC010	200343125	04023		Position	True Pos	0.115	0.015	
3	VOLS-1024-44DAC010	200343125	04032		Position	17 pos Tru	0.387	0.057	
4	VOLS-1024-44DAC010	200343125	04118		Position	True Pos	0.11	0.01	
5	VOLS-1024-44DAC010	200343125	08008		Position	4 pos Tru	0.417	0.387	
6	VOLS-1024-44DAC010	200343125	08026		Profile of	Profile of	0.51 (Min	0.26	
7	VOLS-1024-44DAC010	200343125	10003		Profile of	Profile of	0.51 (Min	0.26	
8	VOLS-1024-44DAC010	200343125	10004		Profile of	Profile of	0.887 (Min	0.537	
9	VOLS-1024-44DAC010	200343125	10003		Position	12 pos Tru	0.487	0.157	
10	VOLS-1024-44DAC010	200343125	10035		Profile of	4 Pos Prof	0.200 (-	0.150	
11	VOLS-1024-44DAC010	200343125	10039		Position	34 Pos Tru	0.639	0.386	
12	VOLS-1024-44DAC010	200343125	10041		Profile of	Profile of	0.259 (Min	0.009	
13	VOLS-1024-44DAC010	200343125	10042		Profile of	Profile of	0.767 (Min	0.637	
14	VOLS-1024-44DAC010	200343125	10052		Profile of	4 Pos Prof	0.240 (-	0.140	
15	VOLS-1024-44DAC010	200343125	11024		Profile of	Profile of	0.356 (Min	0.096	
16	VOLS-1024-44DAC010	200343125	00010		Flatness	All Tolerances apply when measured in restrained condition. With Diam B restrained within 0.05 [0.002] and surface A held to 0.05 [0.002]. Flatness Measured at 0.088mm or 0.038 out of tolerance Please see attached 44DAC01070 support pack			
17									
18									
19									
20									
21									
22									
23									
24									
25									
26									
27									
28									
29									
30									
31									
32									
33									
34									
35									
36									
37									
38									
39									

Project 2: Color Codes Extraction from Images for the Fokker Plant

The final output of the script is a set of Excel files, each corresponding to a PDF file processed during the operation. Each Excel file contains color codes as column headers and the associated numbers as values. The generated Excel files can be further analyzed and utilized for various purposes, such as data visualization and insights.

GUI



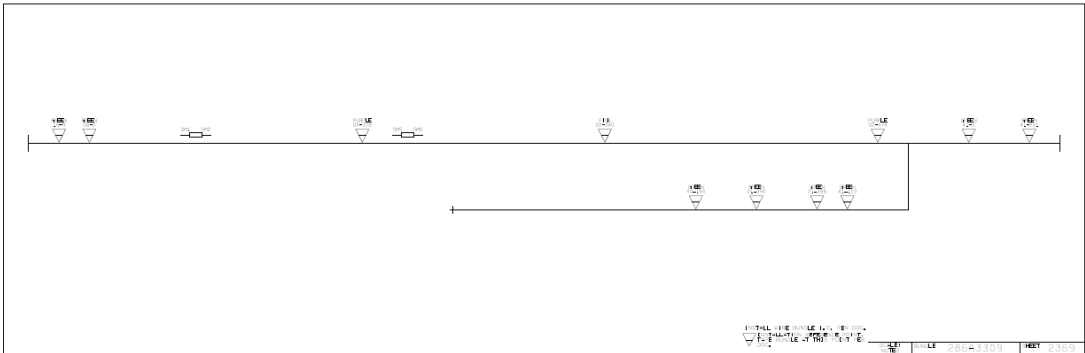
tk

Enter main folder path:

Enter temp folder path:

Enter output folder path:

Input given by user (Sample pdf)



Final Output Generated

File Home Insert Page Layout Formulas Data Review						
Clipboard Font Paragraph Styles						
A1 GREEN						
1	GREEN	PURPLE	PINK			
2	15-9	04-378	02-340			
3	06-2	02-474				
4	21-288					
5	41-123					
6	41-2					
7	1-2					
8	41-221					
9	40-144					
10	21-290					
11						
12						

Project 3: Generalized Data Mining from PDF and Word Documents

The tool is designed to extract various types of data from PDF and Word documents, including tables, equations, relevant text based on keywords, figure captions, table captions, and images. The tool also includes a GUI (Graphical User Interface) using PySimpleGUI, making it easy for users to interact with and use the features.

GUI

The screenshot displays the 'Data Extractor' application window. It features a light purple background and a standard Windows-style title bar with minimize, maximize, and close buttons. The interface is organized into several sections, each with a label and input fields or buttons. At the top, there's a section for 'Thousand & Decimal separator in Word document:' with two radio buttons. Below this are fields for 'Source for Word document:', 'Source for PDF document:', and 'Source of Excel file:' with corresponding 'Browse' buttons. A section titled 'Choose any of the following options:' contains three unchecked checkboxes: 'Apply cell border (All Borders)', 'Copy text format (Bold, Italic & Underline)', and 'Convert cell "-" to 0'. Further down, there are fields for 'Choose folder to save the word tables:', 'Enter file name (Ex: file.xlsx):', 'Choose folder to save the pdf tables:', and 'Enter file name (Ex: file.xlsx):', each with a 'Browse' button. A series of buttons are arranged vertically: 'Extract Tables', 'Extract Relevant Text', 'Extract Equations', 'Extract Figure Captions', 'Extract Table captions', and 'Extract Images'. Each of these buttons is preceded by a label and a 'Browse' button. At the bottom, there's a large empty text area and an 'Exit' button.

Data Extractor

Thousand & Decimal separator in Word document:
☒ 1,234,567,999 ☐ 1.234,567.999

Source for Word document:

Source for PDF document:

Source of Excel file: ☐ Create a blank Excel file

Choose any of the following options:

☐ Apply cell border (All Borders)

☐ Copy text format (Bold, Italic & Underline)

☐ Convert cell "-" to 0

Choose folder to save the word tables:

Enter file name (Ex: file.xlsx):

Choose folder to save the pdf tables:

Enter file name (Ex: file.xlsx):

Source for PDF or Word Document (PDF or DOCX):

Enter the word(s) or phrase(s)(separated by a semicolon):

Choose file to save relevant text(DOCX file):

Choose file to save equations (DOCX file):

Choose file to save the figure captions (DOCX file):

Choose file to save the table captions (DOCX file):

Choose folder to save the extracted images:

Output Generated:

1. Tables Extracted:

RESTRAINT SUMMARY REPORT: Loads On Restraints									
A	B	C	D	E	F	G	H	I	J
1	RESTRAIN SUMMARY REPORT: Loads On Restraints								
2	Various Load Cases								
3	LOAD CASE DEFINITION KEY								
4	CASE(HYD) WW+HP								
5	CASE (OPE) W+T1+P1								
6	CASE (SUS) W+P1								
7	CASE (EXP) L6=L2-L4								
8	NODE 10								
9	FX N.m	FY N.m	FZ N.m	MX N.m	MY N.m	Mz N.m			
10	(HYD)	26	-190	0	0	0	-209		
11	(OPE)	70	-574	0	0	0	-6604		
12	(SUS)	83	-563	0	0	0	-7055		
13	(EXP)	-14	-11	0	0	0	451		
14	MAX	83/3	-574/3	0	0	0	-705.5/3		
15	Rigid ANC								
16	NODE:15	FX N.m	FY N.m	FZ N.m	MX N.m	MY N.m	Mz N.m		
17	(HYD)	-26	-336	0	0	0	902		
18	(OPE)	-70	-935	0	0	0	9671		
19	(SUS)	-83	-947	0	0	0	1749		
20	(EXP)	14	11	0	0	0	-78		
21	MAX	-83/4	-947/3	0	0	0	583		

2. Equations Extracted into a Word Document:

The screenshot shows the Microsoft Word interface. The title bar at the top reads "equations.docx [Compatibility Mode] - Word". The ribbon is set to the "View" tab, with the "Paragraph" group selected. The ribbon shows various icons for paragraph formatting, including bullet points, indentation, and alignment. Below the ribbon, the document content is displayed. It starts with a paragraph "The Equations present in the file are:" followed by four mathematical equations, each on a new line and preceded by a tab character. The equations are: $r = St = MT/2Z$, $Sb = \sqrt{(IIMi) 2 + (IoMo) 2/Z}$, $E = \sqrt{Sb2 + 4St2}$, and $CASE (EXP) L6 = L2 - L4$.

equations.docx [Compatibility Mode] - Word

View Tell me what you want to do...

Paragraph Styles

The Equations present in the file are:

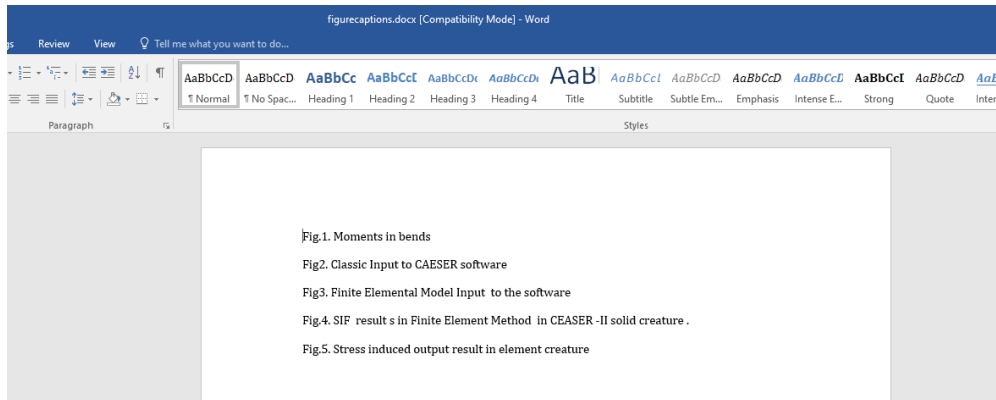
$$r = St = MT/2Z$$

$$Sb = \sqrt{(IIMi) 2 + (IoMo) 2/Z}$$

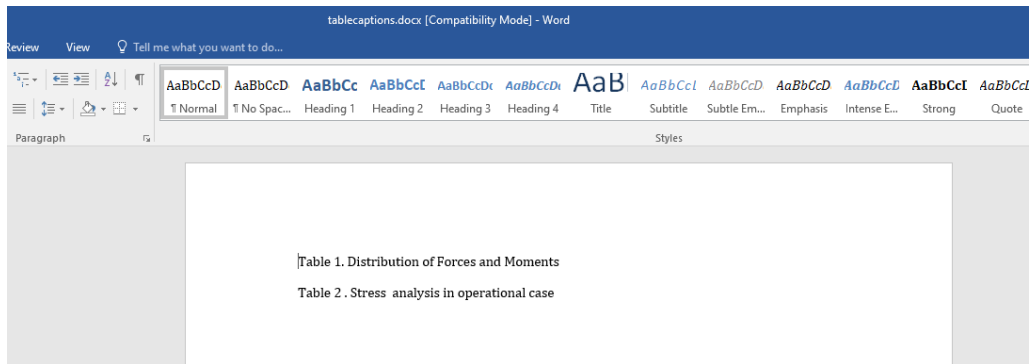
$$E = \sqrt{Sb2 + 4St2}$$

$$CASE (EXP) L6 = L2 - L4$$

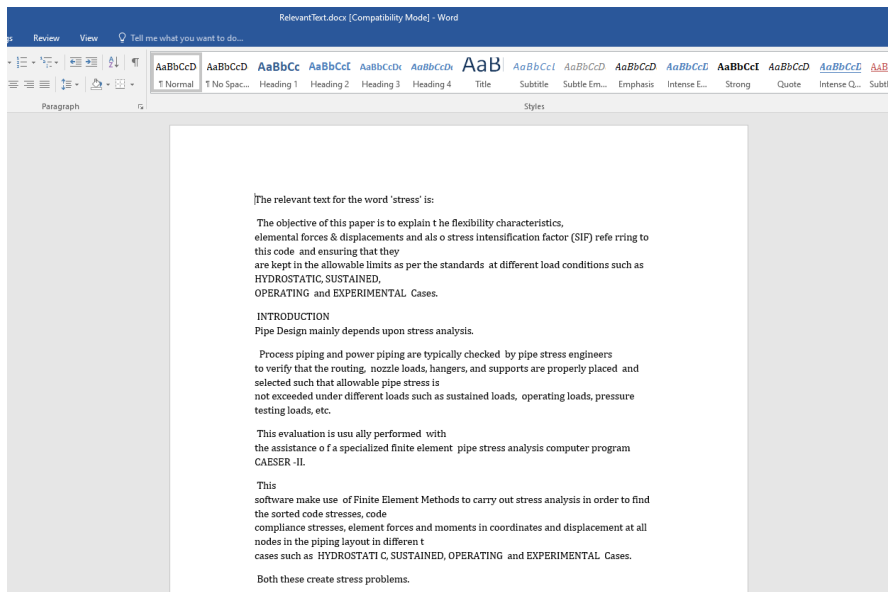
3. Figure Captions Extracted into a Word Document:



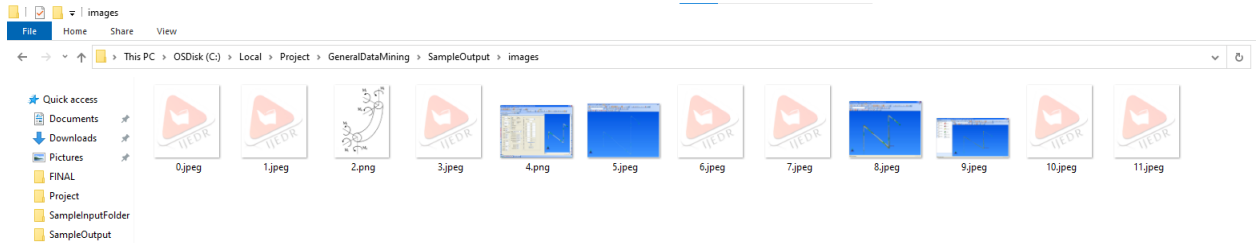
4. Table Captions Extracted into a Word Document:



5. Relevant text based on user input Extracted into a Document:



6. All Images present in the pdf/word Extracted into a Folder:



8. Testing and Performance

All the Python script were tested on various sample files to ensure accurate data extraction and organization. The performance of the script was evaluated on a range of input file sizes to assess its efficiency and handling of larger volumes of unstructured data, image data as well as pdf and word documents.

9. Conclusion

The successful implementation of Data Extraction from Non-Conformance Reports (NCRs) project allowed for the efficient extraction of specific data from Non-Conformance Reports (NCRs) and the creation of a structured Excel file for further analysis. The automation of this data extraction process significantly reduces manual efforts and enhances data accessibility and usability, benefiting the SPE Team and the wider organization.

The successful implementation of the color code extraction project allowed for the efficient conversion of PDF files to images and the extraction of color codes and associated numbers from the images. The utilization of parallel processing using threading improved the overall performance and scalability of the solution. The generated Excel files provide valuable datasets on various products, enabling the GKN Aerospace team to gain useful insights for decision-making and process improvements.

The successful implementation of the Generalized Data Mining from PDF and Word Documents project enabled the extraction of various types of data from PDF and Word documents, including tables, equations, relevant text based on keywords, figure captions, table captions, and images which can be used for multiple purposes. The tool also includes a GUI (Graphical User Interface) using PySimpleGUI, making it easy for users to interact with and use the features.

10. Future Scope

10.1 Non-Conformance Reports (NCRs) Data Extraction:

Advanced Analytics: The curated dataset of extracted data from NCRs can be used for advanced analytics and data-driven decision-making. By employing machine learning algorithms and statistical analysis, the company can identify patterns, root causes of non-conformance, and potential areas for process improvement.

Real-time Monitoring: Integrating the automated data extraction process with real-time NCR generation systems can enable the continuous monitoring of non-conformance instances. This proactive approach allows the company to take prompt corrective actions and improve quality.

10.2 Color Code Extraction from Images:

Expanded Applications: The algorithms developed for color code extraction can be extended to analyze images from different sources, such as visual inspections, quality control checks, and product testing. This opens up opportunities to use color information in various contexts throughout the company's operations.

Integration with Computer Vision Systems: Integrating the color code extraction algorithms with computer vision systems can enhance automation in visual inspection processes, ensuring faster and more reliable detection of color defects.

10.3 Generalized Data Mining from PDF and Word Documents:

Integration with Data Management Systems: Integrating the data extraction algorithms with the company's data management systems can streamline data entry processes and ensure that valuable information from documents is readily available for analysis.

Support for Additional Document Formats: Expanding the algorithms to support additional document formats beyond PDF and Word, such as PowerPoint presentations and web pages, can broaden the scope of data extraction and analysis.

Automation and Process Optimization: By using these algorithms to automate manual data extraction processes, the company can optimize workflows, reduce manual errors, and free up valuable human resources for higher-value tasks